Машинный перевод без параллельного текста

Мазуров Михаил

October 2018

1. Abstract

Рассматривается задача машинного перевода с одного языка на другой. Подготовка корпуса параллельных предложений является ресурсоемкой задачей. Предлагается метод построения нейросетевую модель, которая сможет переводить фразы из одного языка в другой и наоборот. Метод основан на модели Seq2Seq. Предлагается отображать предложения из двух языков в общее векторное пространство. Вычислительный эксперимент проводится на паре языков "русский-украинский".

2. Введение

Эта статья посвящена проблеме машинного перевода из пары языков без параллельных пар предложений, так как для исследовательских целей найти большой набор размеченных данных сложно, но можно создать вероятностную модель перевода статей из открытых источников. Основной идеей подхода является оптимизация seq2seq модели при сопоставлении векторных пространств слов на паре языков "русский-украинский".
Для тренировки перевода будут использованы рекурентные нейронные сети для кодировки и декодирвки с векторным пространством в нужный нам язык, так как на небольших данных этот метод позволяет использовать предыдущий "опыт"модели. Этот подход является многообещающим, он был протестирован на паре языков "английский - французский показав хорошие результаты на уровне 27 BLEU (Bahdanau, Cho, & Bengio, 2014)
Также будут использоваться attention-механизмы, которые показали свою состоятельность, давая улучшение в исследованиях порядка 5 BLEU (Luong, Pham, & Manning, 2015)
Для борьбы с переобучением кодировщиков будет использоваться механизм зашумления предложений, чтобы кодировщики не учили переводы предложений слово-в-слово. Предлагается использовать автокодировщик, который будет учиться убирать шум в предложениях без перевода на другой язык.(Kim & Ney, 2018)

3. Постановка задачи

Мы работаем с двумя выборками на разных языках: $\mathfrak{D}^{\text{src}} = [\mathbf{s}_1^{\text{src}}, \dots, \mathbf{s}_{m_{\text{src}}}^{\text{src}}]$.
и

- Обучающая выборка на первом языке: $\mathfrak{D}^{\text{src}} = [\mathbf{s}_1^{\text{src}}, \dots, \mathbf{s}_{m_{\text{src}}}^{\text{src}}]$.
- Обучающая выборка на втором языке: $\mathfrak{D}^{\text{tgt}} = [\mathbf{s}_1^{\text{tgt}}, \dots, \mathbf{s}_{m_{\text{tgt}}}^{\text{tgt}}]$.
- Максимальная длина предложения: l.
- Мощность словаря на первом языке: V^{src}.
- Мощность словаря на втором языке: V^{tgt}.
- Предложение на первом языке: $\mathbf{s}^{\text{src}} = [x_1, \dots, x_l], \quad x_i \in \{1, \dots, V^{\text{src}}\}$.
- Предложение на втором языке: $\mathbf{s}^{\text{tgt}} = [y_1, \dots, y_l], \quad y_i \in \{1, \dots, V^{\text{tgt}}\}$.
- Энкодер: $\mathbf{f}$. (Если потребуется — ставьте индекс src или tgt около модели).
- Декодер: $\mathbf{g}$. (Если потребуется — ставьте индекс src или tgt около модели).
- Скрытое представление: $\mathbf{h}$.
- Валидационная выборка: $\mathfrak{D}^{valid} = \{(\mathbf{s}_1^{\text{src}}, \mathbf{s}_1^{\text{tgt}}), \dots, (\mathbf{s}_{m_{\text{valid}}}, \mathbf{s}_{\text{valid}})\}$.
- Параметры Seq2Seq-модели: $\mathbf{w}$.

Список литературы

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kim, Y., & Ney, J. G. H. (2018). Improving unsupervised word-by-word translation with language model and denoising autoencoder.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.